A Comment on the ROC Curve and the Area Under it as Performance Measures

Caren Marzban*

Center for Analysis and Prediction of Storms
University of Oklahoma, Norman, OK 73019

and

Department of Statistics
University of Washington, Seattle, WA 98195

April 16, 2003

*http://www.nhn.ou.edu/~marzban

Abstract

The Receiver Operating Characteristic (ROC) curve is a two dimensional measure of classification performance. The area under the ROC curve (AUC) is a scalar measure gauging one facet of performance. In this note, an attempt is made to relate the shape of the ROC curve (and the area under it) to features of the underlying distribution of forecasts, allowing for an interpretation of the former in terms of the latter. To that end, three idealized examples are considered analytically as models of the underlying distribution of forecasts. The examples cover non-probabilistic and probabilistic forecasts. The exact expressions for ROC and AUC are derived, in turn, allowing for their interpretation in terms of features of the underlying distributions. For example, a asymmetric ROC curve can be interpreted as unequal variances for the distributions. Furthermore, it is shown that AUC discriminates well between "good" and "bad" models, but not between "good" models.

1 Introduction

In the words of the late Alan Murphy “Performance is a multifaceted thing.” [1] However, that fact can often be obfuscated by technical difficulties. For example, in regression and classification problems, it is relatively straightforward to select a single scalar (i.e., 1-dimensional) measure such as mean squared error, and proceed to optimize it. Similarly, in assessing the superiority of one model over another, it is standard practice to compare the respective values of a single scalar measure like chi-squared. A number of such scalar measures common in meteorological circles are examined in Marzban (1998). However, if one adopts several measures to optimize simultaneously, then the optima are often not unique.[2]

The adoption of a scalar measure presumes that only some specific aspect of performance is of importance in the corresponding problem. There exist situations wherein that unique facet can be identified, but that is not the case in most realistic problems. In general, a proper assessment of the performance of some model or forecasting scheme is an extremely difficult task mostly due to the multidimensional nature of performance itself. It is quite likely that one model will outperform another model in terms of one scalar measure of performance, but not in terms of another measure. This is not a defect of the models, but a simple consequence of the fact that performance is a multidimensional quantity.

For probabilistic forecasting a number of multidimensional tools have been developed in meteorological circles (Murphy and Winkler 1987, 1992). Naturally, most are diagrams, i.e. 2-dimensional. As such, they can provide a relatively complete assessment of performance, at least for the 2-class problem wherein one forecasts the probability of belonging to one of two classes, e.g., tornado/nontornado, rain/no-rain, etc. Larger-class problems can often be treated in terms of several 2-class problems, albeit with some loss of information.

[1] Personal communication.

[2] The task of simultaneously optimizing several criteria belongs to the topic of multiobjective (or Pareto) optimization (Miettinen, 1999).

For nonprobabilistic or categorical forecasts, the contingency table (Wilks 1995) provides a complete representation of performance for any number of classes, but it is difficult to display and interpret. However, the 2-class problem is unique in that the performance of a model can be displayed in a (2-dimensional) diagram. One such diagram is the Receiver Operating Characteristic (ROC) curve. Its history is rich and lively; an exhaustive list of references has been compiled by K. Zou (2001). It is a graphic representation of the performance of a model that produces categorical forecasts of a 2-class event. Specifically, it is a parametric plot of the hit rate (or probability of detection) vs. the false alarm rate, as a decision threshold is varied across the full range of the forecast quantity. Fig. 1 displays three ROC curves representing different levels of performance (from Marzban and Witt 2001). The diagonal line corresponds to random forecasts (i.e., poor performance), and performance is improved the further the curve bows away from the diagonal. In fact, the area under the ROC curve (AUC) is often taken as a scalar measure of performance. An area of 0.5 would reflect random forecasts, while an area of 1 suggests perfect forecasts. Distilling a 2-dimensional entity like the ROC curve to a scalar quantity like AUC implies that the latter must depend on a unique combination of the parameters of the underlying distribution. This affinity between a measure of performance and some specific combination of the underlying parameters will be made explicit in the following. In this figure, error bars are also displayed for the purpose of assessing the statistical significance of the performance.

The ROC curve has great utility in assessing performance in a multi-dimensional (or non-scalar) setting. The more fundamental quantity, however, is the class-conditional distribution of the forecasts, i.e., the distribution of the forecast quantity for each class, separately. After all, the quantities from which an ROC curve is derived - hit rate and false alarm rate - are areas under these distributions, above or below some decision threshold (Masters 1993). Although the computation of ROC curves does not require knowledge of these distributions, it is natural to utilize the connection between the ROC curve and the underlying distributions to infer something about the latter. Here, several characteristic features of ROC curves will be identified with features of the underlying distributions. As such, the shape of the ROC curve can be interpreted. In addition to placing the assessment of performance on a more fundamental foundation, knowledge of the underlying distributions can guide the development of better models. Armed with the connection between the ROC curve and the underlying distributions, one can also further interpret the meaning of AUC.

2 Preliminaries

Henceforth, the two classes are labeled as 0 (nonevent) or 1 (event). The 4 elements of the 2×2 table are C_1 (C_4), the number of correctly classified nonevents (events), and C_2 (C_3), the number of incorrectly classified nonevents (events). Although the table has 4 elements, there are only two degree of freedom if the class-conditional sample sizes are fixed, because $N_0 = C_1 + C_2$, and $N_1 = C_3 + C_4$. The contingency table, in turn, can be reduced to a host of scalar measures of performance, but in order to preclude any loss of information (due to the reduction from two to one degree of freedom) two scalar measures are considered. Two common measures are the hit rate, H, and the false alarm rate, F:

$$H = \frac{C_4}{C_3 + C_4} \quad , \quad F = \frac{C_2}{C_1 + C_2} \quad .$$

The ROC curve is a parametric plot of H vs. F as a decision threshold is varied across the full range of the forecast quantity. Although the amount of bowing away from the diagonal is a measure of performance, the specific shape (or features) of the curve are informative, and yet often ignored. That shape is determined by the class-conditional distributions of the forecast quantity. Therefore, it is possible to translate the shape of an ROC curve to some information regarding the underlying distributions.

In order to get a handle on the specific relation between the shape of the ROC curve and the underlying conditional distributions, it is instructive to consider some models. For example, one can model the class conditional distributions with gaussians, with means μ_0, μ_1, and standard deviations σ_0, σ_1. Then it is possible to write down analytic expression for H and F. In particular, if $\sigma_0^2 > \sigma_1^2$, then

$$H = \frac{1}{2}[1 - erf(\frac{t-\mu_1}{\sqrt{2}\sigma_1})] \ \ , \ \ F = \frac{1}{2}[1 - erf(\frac{t-\mu_0}{\sqrt{2}\sigma_0})],$$

where t is the decision threshold, and $erf()$ is the Gaussian Error Function. Exchanging $0 \leftrightarrow 1$ in these equations leads to the respective rates for the case $\sigma_1^2 > \sigma_0^2$. Given the appearance of $erf()$, it is difficult to arrive at an explicit functional expression for the ROC curve. For this reason, two different approximations will be considered here; uniform distributions (Fig. 2a) and bell-shaped distributions (Fig. 3a). Furthermore, when the forecast quantity is a probability, then gaussian approximations are inadequate, since probabilities are restricted to the range 0 to 1, while gaussians are unbound. For this situation, the distributions will be modeled as simple approximations shown in Fig. 4a. These overly simplified models suffice in arriving at some general conclusions.

3 Uniform

A generic situation involving forecasts with uniform distributions is shown in Fig. 2a. There are four parameters involved - two means, μ_0 and μ_1, and two "standard deviations", σ_0 and σ_1. Without loss of generality, it is assumed that $\mu_1 \geq \mu_0$. It is then straight forward to show that the false alarm rate and the hit rate are given by

$$F = \frac{\mu_0 + \sigma_0 - t}{2\sigma_0}, \ \ H = \frac{\mu_1 + \sigma_1 - t}{2\sigma_1}, \tag{1}$$

where t is the decision threshold above (below) which a case is classified into class 1 (0).

The ROC curve follows immediately from (1):

$$H = \frac{\sigma_0}{\sigma_1} \ F \ + \ \frac{\delta\mu + \delta\sigma}{2\sigma_1} \ \ , \tag{2}$$

where $\delta\mu = \mu_1 - \mu_0$ and $\delta\sigma = \sigma_1 - \sigma_0$. Fig. 2b displays the situation. It can be seen that the ROC curve consists of three line segments, with the equation for the middle segment given by (2).

Several observations can be made. First, (2) implies that two models with different means and standard deviations can yield the same ROC curve if they have the same slope and intercept (see Fig. 2b). As such, the ROC curve does not uniquely specify the underlying parameters. In other words, there is a family of underlying distributions that give rise to the same ROC curve.

Second, the length of the vertical segment is determined by two quantities, $\delta\mu$ and σ_0/σ_1. This is sensible since the goodness of the underlying model is determined by both quantities. By contrast, the slope of the middle segment depends only on the ratio of the standard deviations (and not $\delta\mu$). As such, the inequality of σ_0 and σ_1 reflects itself as an ROC line that is not parallel to the diagonal.

Given the analytic expression for the ROC curve (2), it is then possible to compute the area under the curve, AUC:

$$AUC = 1 - \frac{1}{8}\left(\frac{\Delta}{\sqrt{\sigma_0\sigma_1}}\right)^2 \tag{3}$$

where

$$\Delta = \delta\mu - (\sigma_0 + \sigma_1). \tag{4}$$

This is an instance of what was previously referred to as the affinity between a measure of performance (like AUC) and a certain function of the underlying distribution parameters. In this case, AUC has an affinity for the combination $\Delta = \delta\mu - (\sigma_0 + \sigma_1)$. This is easy to understand: a better model should have a larger AUC, which in turn means that it should have a smaller Δ. And that means a larger $\delta\mu$, and a smaller $\sigma_0 + \sigma_1$. In short, model selection based on AUC selects for sharp and widely separated class-conditional distributions, where sharpness is measured by the sum of the standard deviations, and separation is gauged by the difference in their means.

Moreover, as a function of Δ, AUC is a parabola. Fig. 2c shows an instance for $\sigma_0 = \sigma_1 = 0.05$ and $\sigma_0 = 0.05, \sigma_1 = 0.10$. It is this parabolic behavior that explains the often experienced empirical finding that AUC discriminates well between good and bad models, but not between good models. Note that a perfect model corresponds to $\Delta = 0$ in Fig. 2c where the curves flatten off. Therefore, good models have approximately the same values of AUC.

Equation (4) also implies that any set of models whose μ and σ values lead to the same value of Δ, also yield the same value for AUC. In other words, a given value of AUC corresponds to a range of values for $\mu_0, \mu_1, \sigma_0, \sigma_1$ that reside on the plane defined by $\mu_1 - \mu_0 - (\sigma_0 + \sigma_1) =$ constant. Note that this plane is a larger space than the space of parameters that yield equivalent ROC curves. The former is the plane given by $\mu_1 - \mu_0 - (\sigma_0 + \sigma_1) =$ constant, while the latter is given by the intersection of two planes (i.e. a line) - the plane defined by $\sigma_0/\sigma_1 =$ constant and that defined by $(\delta\mu + \delta\sigma)/2\sigma_1$ =constant; see (2). In short, AUC is even more ambiguous in selecting an underlying model than the ROC curve itself.

4 Bell-shaped

Most distributions arising in practice are not uniform but bell-shaped. One approximation to bell-shaped distributions is shown in Fig. 3a. [3] The added complexity is that no single expression for F and H can be written. However, it is possible to write analytic expressions in each of the regions: $\mu_0 - \sigma_0 \leq t \leq \mu_0$, $\mu_0 \leq t \leq \mu_1 - \sigma_1$, and $\mu_1 - \sigma_1 \leq t \leq \mu_0 + \sigma_0$. The remaining region to the right has $F = 0$. For example, in the second and third regions one has:

$$F = \frac{1}{2}\left(\frac{\mu_0 + \sigma_0 - t}{\sigma_0}\right)^2, \tag{5}$$

and the corresponding H's are

$$H = 1 \quad \text{and} \quad H = 1 - \frac{1}{2}\left(\frac{t - \mu_1 + \sigma_1}{\sigma_1}\right)^2 \quad , \tag{6}$$

respectively. The ROC curves corresponding to the three regions are shown in Fig. 3b. Unlike the case of uniform distribution, the ROC curve for bell-shaped distributions is more realistic, with its characteristic bowing above the diagonal line. A Common feature, however, is that the quickness with which the curve rises (i.e., the vertical segment in Fig. 3b) is determined by σ_0/σ_1, and $\delta\mu$.

From the endpoins of the middle segment (Fig. 3b), one can conclude that if the empirical ROC curve bows up in a non-symmetric fashion, then one can interpret

[3] Any similarity with membership functions in Fuzzy Sets is coincidental.

that as $\sigma_0 \neq \sigma_1$; a symmetric ROC curve implies $\sigma_0 = \sigma_1$. Specifically, if the bowing is mostly to the left, then $\sigma_0 < \sigma_1$. Bowing to the right suggests $\sigma_0 > \sigma_1$.

Also, the two extremes of the curves - $F = 0$ and $H = 1$ - convey some useful information as well. Note that if $\delta\sigma = \delta\mu$, then the right extreme of the curve meets the (1,1) point without overlapping the $H = 1$ line. Similarly, $\delta\sigma = -\delta\mu$ implies that the left extreme of the curve meets the (0,0) point without overlapping the $F = 0$ axis. Therefore, the amount of overlap of the curve and the two axes is a measure of the distance between the two means *relative to* the difference between the standard deviations.

AUC can be computed to be

$$AUC = 1 - \frac{1}{8}\left(\frac{\Delta}{\sqrt{\sigma_0\sigma_1}}\right)^4 \tag{7}$$

This equation confirms that the AUC has an affinity for the quantity Δ defined in (4). Furthermore, noting the quartic power of Δ, in contrast to the quadratic power in (3), it is clear that this AUC is highly nonlinear. Fig. 3c displays this quartic dependence. Comparing Fig. 3c with Fig. 2c, it becomes clear that bell-shaped distributions lead to more nonlinearity in AUC than uniform distributions. As such, empirical AUC curves are apt to be highly nonlinear. This further flattening of the AUC curve exacerbates AUC's inability to discriminate between good models.

5 Probabilistic

In some situations the forecast quantity is probabilistic. Then, the previous approximations are invalid because the forecast quantity is between 0 and 1, while the previous two examples assume an unbound forecast quantity. Fig. 4a shows the type of approximation that will be examined here for probabilistic forecasts. Note that in this approximation, the underlying distributions have no associated standard deviation; the only parameters are the two "means", μ_0 and μ_1.

Three different regions must be considered: $t \leq \mu_0, \mu_0 \leq t \leq \mu_1$, and $t \geq \mu_1$. Unlike the previous examples, here there exists no region in which either H or F are

1. The functional dependence of H on F (i.e., the ROC curve) in each of the three regions is

$$H = 1 - \frac{\mu_0}{\mu_1}(1-F) \ \ , \ \ H = 1 - \frac{1}{\mu_1}[1 - \sqrt{(1-\mu_0)\ F}\ ]^2 \ \ , \ \ H = (\frac{1-\mu_0}{1-\mu_1})\ F \ \ , \quad (8)$$

respectively. Note that the ROC curve for first and third regions are linear, while that of the middle section is not. Fig. 4b displays the ROC curve for some values of the parameters. The values of H and F at the boundaries between the regions are also shown.

It is worth pointing out that the probabilistic case is different from the previous examples in one important way. Whereas in the previous examples ROC was shown to have an affinity for a given combination of the underlying parameters, effectively reducing the dimensionality of the problem by one, in the probabilistic case, both parameters - μ_0, μ_1 - independently affect the ROC curve. This will be made more clear, below. For this reason, the shape of the ROC curve is not as readily interpretable as it is in the previous examples.[4]

In spite of this slight complication, Fig. 4b still offers some useful information. For instance, consider the slopes of the linear segments. From (8) these are μ_0/μ_1 and $(1-\mu_0)/(1-\mu_1)$, for the top and bottom segments, respectively. First, note that a large slope in one is accompanied by a small slope in the other. In fact, the relationship between the slopes is linear, and so, the behavior of the extreme ends of the ROC curve are linearly tied.[5] Moreover, a large slope for the bottom segment translates to $\mu_0 << \mu_1$, and vice versa. By contrast, a large slope for the top segment corresponds to $\mu_0 >> \mu_1$.

The point along the middle segment at which the slope is 1 can be easily computed

[4]The dependence of ROC on both parameters is not too surprising. If the forecast quantity were unbound (as in the previous two examples), then one would expect the difference between the means $(\mu_0 - \mu_1)$ to play a central role. And, as shown, it does in those examples. However, in the probabilistic case, given the finite bound, one might equally consider the product of the means as important. In fact, as shown in the expression for AUC, both combinations play equally important roles.

[5]The linear relationship is $\mu_1(slope_{top}) + (1-\mu_1)(slope_{bottom}) = 1$. It is likely that in a better approximation than those shown in Fig. 4a, the linearity of the relation will be lost. However, the behavior of the extremes of the ROC curve will still be tied. This is a general consequence of the finite bound of probabilistic forecasts.

from (8). This point is important in that it suggests the corresponding value of the threshold as an "optimal" value. Its coordinates (F_c, H_c) are given by

$$F_c = \frac{1-\mu_0}{(\mu_1-\mu_0+1)^2} \quad , \quad H_c = 1 - \frac{\mu_1}{(\mu_1-\mu_0+1)^2} \quad . \tag{9}$$

It follows from the expression for F_c that the ROC curve will bow to the left if $\mu_0 \sim 0.5$, and to the right if $\mu_1 \sim 0.5$.

Finally, the AUC can be computed to be

$$\text{AUC} = \frac{1}{2} + \frac{1}{2}(\mu_1-\mu_0) - \frac{(\mu_1-\mu_0)^3}{3\mu_1(1-\mu_0)}. \tag{10}$$

First, note that given the constraints embodied in Fig. 3a, perfect performance (i.e., AUC=1) is never achieved, even for $\mu_0 = 0, \mu_1 = 1$; in that limit, one has AUC=5/6. Second, as anticipated above, AUC depends on two independent quantities - $\mu_1 - \mu_0$ and $\mu_1(1-\mu_0)$. For small values of the former, i.e., low performance, the first two terms in (10) dominate the expression, leading to a linear dependence on $\mu_1 - \mu_0$. However, for higher performance values, the last term begins to penalize (because of the negative sign) AUC in a nonlinear fashion. This nonlinear penalty again leads to a flattening of the AUC, and is another illustration of how AUC does not discriminate well between good models.

Fig. 4c displays AUC as a function of μ_0 and μ_1. The surface is approximately a plane near the x-y plane. As it rises to higher AUC values, it flattens off. This is a consequence of the nonlinearity of AUC. Fig. 4c appears to suggest that the single quantity $\mu_0 + \mu_1$, i.e., the "spine" of the surface, could be the only quantity on which AUC depends. However, it turns out that the surface is also curved horizontally; although difficult to see in the figure, the AUC=constant curves are, in fact, nonlinear curves for larger AUC values. That curvature is due to the "interaction term" $\mu_1(1-\mu_0)$ appearing in (10), which again, is a consequence of probabilistic forecasts.

6 Summary and Conclusion

It is shown that the shape of the ROC curve can be interpreted in terms of the parameters defining the class-conditional distributions of the forecast quantity. Three idealized examples are considered wherein expressions for the ROC curve and the area under the curve (AUC) can be computed analytically. The examples include two instances of forecast quantities - one where the quantity is unbound, ranging from $-\infty$ to $+\infty$, and another bound between 0 and 1, corresponding to probabilistic forecasts.

Conclusions resulting from the derived analytic expressions are numerous. A few of the simpler ones can be summarized as follows:

For non-probabilistic forecasts:

- The asymmetry of the ROC curve can be interpreted as an inequality between the widths, σ_0, and σ_1.
- An ROC curve bowing to the left (and down) suggests $\sigma_0 < \sigma_1$, while a bowing to the right (and up) implies $\sigma_0 > \sigma_1$.
- The amount of overlap of the ROC curve with the x- and y-axis is a measure of the distance between the two means, μ_0, μ_1, relative to the difference between the widths.
- The AUC is good at discriminating between "good" and "bad" models, but not between "good" models.
- The AUC has an affinity for a certain combination of the parameters of the underlying distributions. That combination is of the general form $(\mu_1 - \mu_0) + (\sigma_0 + \sigma_1)$, where μ_i and σ_i are some measure of the central tendency and the width, respectively, of the distribution of forecasts of the i^{th} class.

For probabilistic forecasts:

- ROC curves for probabilistic forecasts are behaviorally "richer" than those of non-probabilistic forecasts, because they depend on all the underlying parameters, independently, without an affinity for a single and specific combination of the parameters.

- ROC curves will bow to the left (and down) if the class-conditional distribution of the 0^{th} class is centered on 0.5. Similarly, a bowing to the right (and up) corresponds to $\mu_1 \sim 0.5$.

- The behavior of the ROC curve at the two ends are tied together. A large slope at the bottom implies $\mu_0 >> \mu_1$, and a large slope at the top suggests $\mu_0 << \mu_1$.

- The AUC is good at discriminating between good and bad models, but not between good models, with this inability inversely proportional to the product $\mu_1(1 - \mu_0)$.

Given the analytic expressions derived for ROC and AUC, it is also possible to compute error-bars or confidence bands for them. It is worthwhile to examine gaussian class-conditional distributions in full complexity. Although, as shown in Section 1, analytic expressions are not readily available, it is possible to examine the problem numerically. This can shed some light on the generality of the results found here. These will be considered in a future article.

Acknowledgement

The author is grateful to Rich Caruana for invaluable discussions and a thorough reading of an early version of this article.

Figure Caption

Figure 1. An example of an ROC diagram with three ROC curves representing different levels of performance quality. The diagonal line corresponds to random forecasts (i.e., poor performance), while the curves bowing away from the diagonal represent higher levels of performance.

Figure 2. The schematics of uniform class-conditional distributions (top), the corresponding ROC curve (middle), and the AUC curve as a function of the quantity $\Delta = \delta\mu - (\sigma_0 + \sigma_1)$.

Figure 3. Same as Fig. 2, but for bell-shaped distributions.

Figure 4. Same as Fig. 2, but for bell-shaped, probabilistic forecasts.

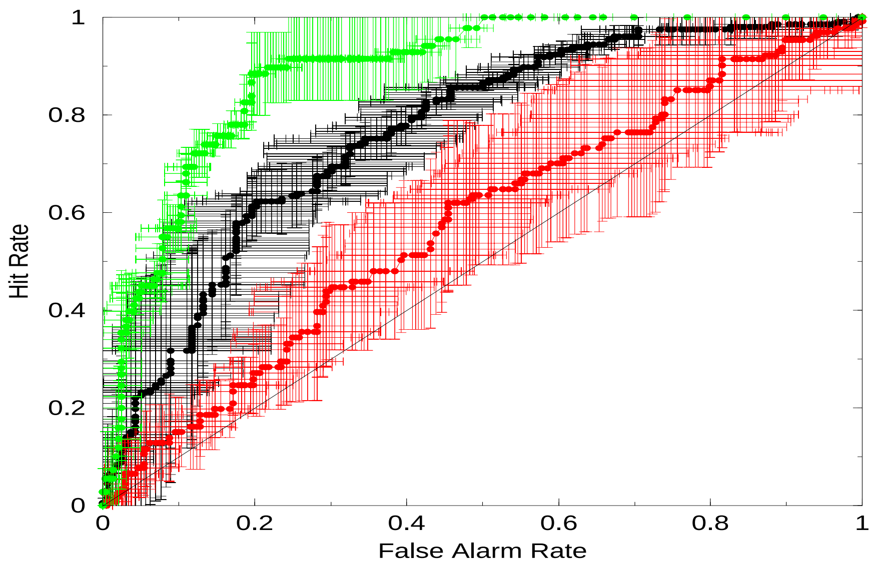
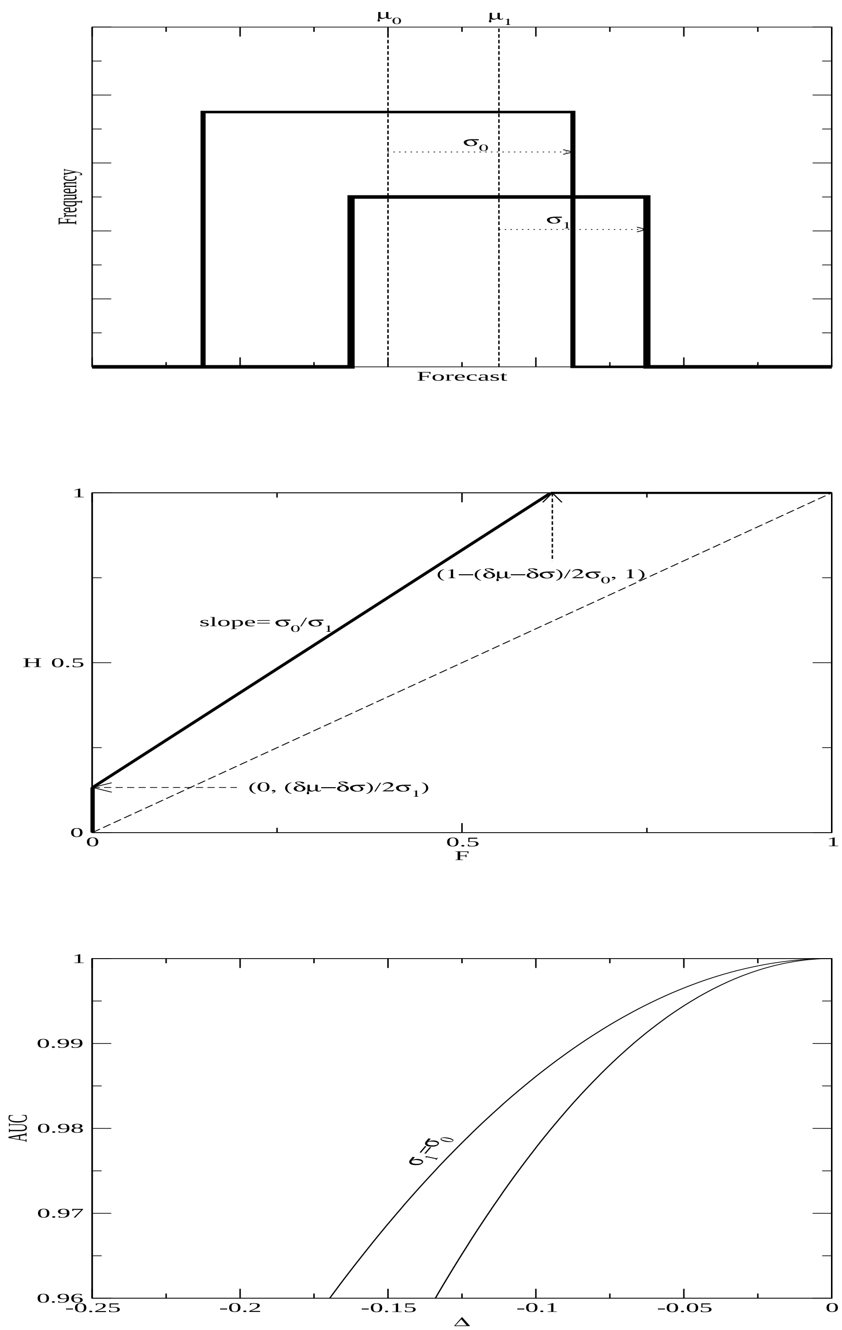
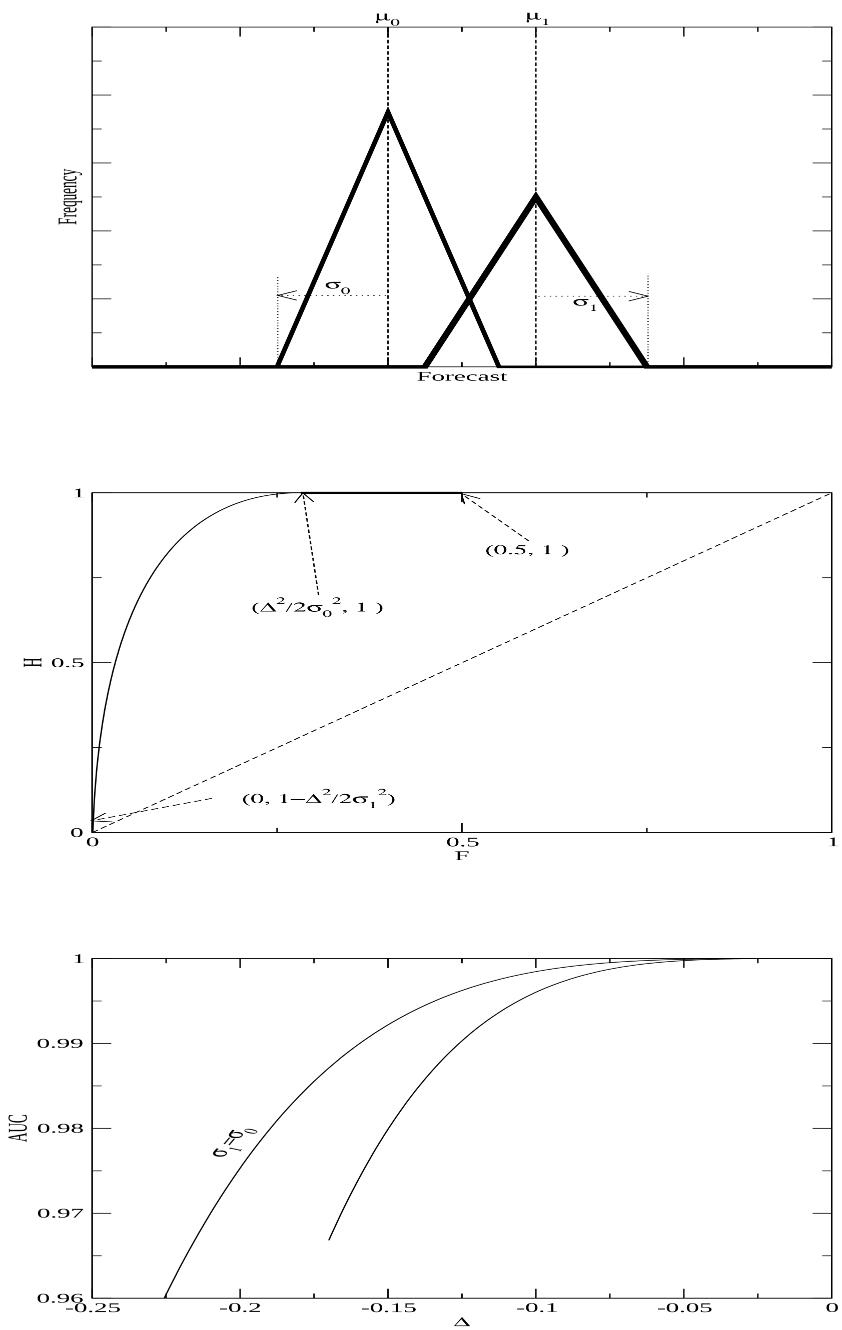


Figure 1.



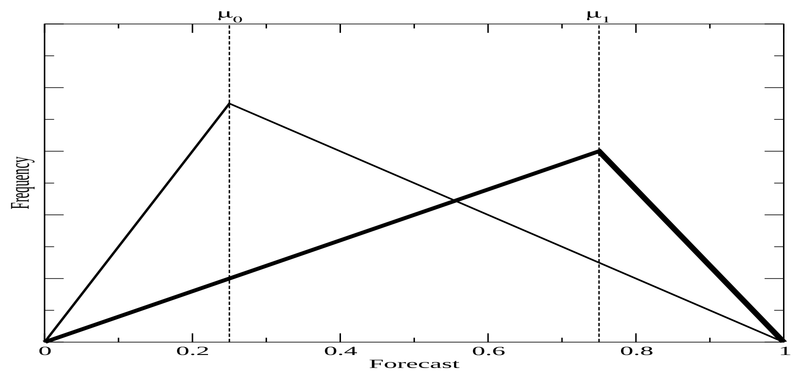
μ0
μ1
σ0
σ1
Frequency
Forecast
1
(1−(δμ−δσ)/2σ0, 1)
slope=σ0/σ1
H 0.5
(0, (δμ−δσ)/2σ1)
0
0
0.5
1
F
1
0.99
AUC
0.98
σ1=σ0
0.97
0.96
−0.25
−0.2
−0.15
−0.1
−0.05
0
Δ

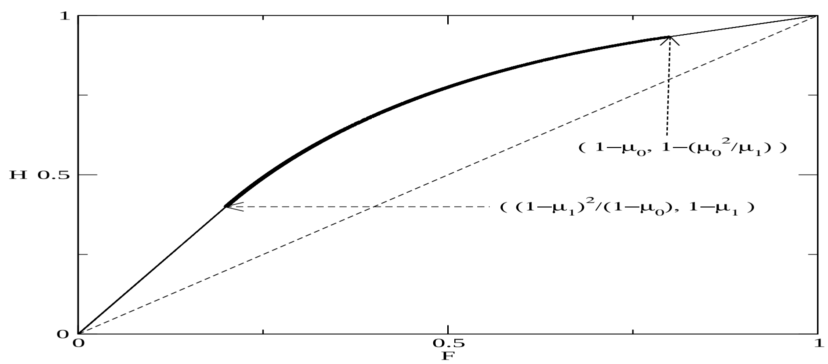
Figure 2.



μ_0
μ_1
Frequency
σ_0
σ_1
Forecast
$(0.5, 1)$
$(\Delta^2/2\sigma_0^2, 1)$
$(0, 1-\Delta^2/2\sigma_1^2)$
H
F
0
0.5
1
AUC
$\sigma_1=\sigma_0$
0.96
0.97
0.98
0.99
1
-0.25
-0.2
-0.15
-0.1
-0.05
0
Δ

Figure 3.


μ_0
μ_1
Frequency
0
0.2
0.4
0.6
0.8
1
Forecast


1
H 0.5
0
0
0.5
1
F
$(1-\mu_0, 1-(\mu_0{}^2/\mu_1))$
$((1-\mu_1)^2/(1-\mu_0), 1-\mu_1)$

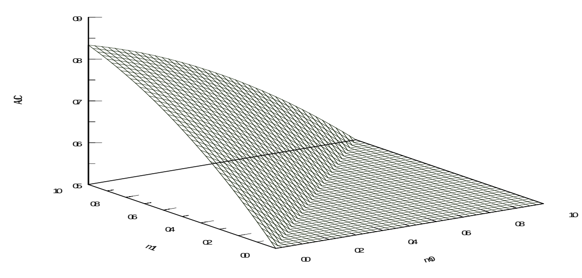

AC

Figure 4.

References

Masters, T., 1993: *Practical neural network recipes in C++.* Academic Press, 493 pp.

Marzban, C. 1998: Scalar measures of performance in rare-event situations, *Wea. Forecasting*, **13**, 753-763.

Marzban, C., and A. Witt, 2001: A Bayesian Neural Network for Hail Size Prediction. *Wea. Forecasting*, **16**, 600-610.

Miettinen, K., 1999: Nonlinear Multiobjective Optimization, Kluwer Academic, Boston, 300 pp.

Murphy, A. H., and R. L. Winkler, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting*, **7**, 435-455.

———, and ———-, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330-1338.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, NY. 467 pp.

Zou, K. H., 2001: http://splweb.bwh.harvard.edu:8000/pages/ppl/zou/roc.html